DOCUMENT RESUME

ED 354 273 TM 019 565

AUTHOR Muthen, Bengt O.
TITLE Advances in Multi-Level Psychometric Models: Latent Variable Modeling of Growth with Missing Data and Multilevel Data. Project 2.6: Analytic Models To Monitor Status and Progress of Learning and Performance and Their Antecedents.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY National Inst. on Alcohol Abuse and Alcoholism (DHHS), Rockville, Md.; Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 92
CONTRACT NIAA-AA-08651-01; R117G10027
NOTE 20p.; Paper presented at the International Conference on Multivariate Analysis (7th, Barcelona, Spain, September 21-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Cluster Analysis; Equations (Mathematics); Individual Differences; Longitudinal Studies; *Mathematical Models; *Multivariate Analysis; *Psychometrics
IDENTIFIERS *Latent Variables; Longitudinal Study of American Youth; *Missing Data; Multilevel Analysis; Structural Modeling

ABSTRACT

Three important methods areas of multivariate analysis that are not always thought of in terms of latent variable constructs, but for which latent variable modeling can be used to great advantage, are discussed. These methods are: (1) random coefficients describing individual differences in growth; (2) unobserved variables corresponding to missing data; and (3) variance components describing data from cluster sampling. An educational achievement dataset of longitudinal observations on secondary mathematics achievement (the National Longitudinal Study of American Youth) is described as a motivating example. It is shown that all three topics can be simply expressed in terms of latent variable modeling that fits into existing and generally available structural modeling software. This approach makes possible a connection between psychometricians and other methodologists interested in latent variable modeling. Interesting extensions of these statistical analyses are discussed. One table presents missing data patterns.

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED354273

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – November 1992

Project 2.6 Analytic Models to Monitor Status and
Progress of Learning and Performance
and Their Antecedents

Analytic Models: Latent Variable Modeling
of Educational Achievement

Advances in Multi-level Psychometric Models: Latent Variable Modeling of Growth With Missing Data and Multilevel Data

Bengt O. Muthén, Project Director
CRESST/University of California, Los Angeles
(310) 206-1226

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532
(310) 825-2917

TM019565

BEST COPY AVAILABLE

The contents of this report were also presented as an invited paper for the Seventh International Conference on Multivariate Analysis, Barcelona, Spain, September 21-24, 1992.

The work reported herein was supported in part by grant AA 08651-01 from NIAAA for the project "Psychometric Advances for Alcohol and Depression Studies" and in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

LATENT VARIABLE MODELING OF GROWTH WITH MISSING DATA AND MULTILEVEL DATA[1]

Bengt Muthen, CRESST/University of California, Los Angeles

1. Introduction

The aim of this paper is to describe three important methods areas of multivariate analysis that are not always thought of in terms of latent variable constructs, but for which latent variable modeling can be used to great advantage: random coefficients describing individual differences in growth; unobserved variables corresponding to missing data; and variance components describing data from cluster sampling. An educational achievement data set will be described as a motivating example. Using the features of the example, it will be shown that all three topics can be simply expressed in terms of latent variable modeling which fits into existing and generally available structural modeling software. This development makes a connection between mainstream statistical methods and work by psychometricians and other methodologists interested in latent variable modeling. Having put the methodology in a general latent variable context, several interesting extensions of the statistical analyses are evident.

2. A General Latent Variable Framework

Analysis of latent variable models is most often carried out by minimizing the following fitting function

$$(1) \qquad \sum_{p=1}^{P} \{ N_p [\ln | \Sigma_p | + \mathrm{tr} (\Sigma_p^{-1} T_p) - \ln | S_p | - r] \} N^{-1} ,$$

where

$$(2) \qquad T_p = S_p + (\bar{y}_p - \mu_p) (\bar{y}_p - \mu_p)' .$$

[1] I thank Ginger Nelson, who provided helpful research assistance.

In maximum-likelihood (ML) estimation of conventional structural equation models with latent variables, this is the fitting function corresponding to independent random samples from P populations with sample sizes N_p and total sample size N. Here, an r-dimensional vector y, say, is observed with sample covariance matrix S_p, sample mean vector $\bar{y}_p$, population covariance matrix Σ_p, and population mean vector μ_p. The terms containing $\ln |S_p| - r$ are offsets so that a perfectly fitting model has the function value of zero. The sample covariance matrices S_p are the ML estimates of the unrestricted Σ_p matrices and are therefore divided by N_p, not $N_p - 1$. Multiplying the minimum value for any model by $2 \times N$ then gives the value of the likelihood-ratio chi-square test of the H_0 model against the H_1 model of unrestricted mean vectors μ_p and covariance matrices Σ_p. Many models do not impose any restrictions on μ_p in which case the second term on the right-hand-side of (2) vanishes and only covariance matrices are involved in the estimation. The simultaneous analysis of several populations is considered when the populations have parameters in common, so that equality constraints of parameters across populations are invoked.

The specification of latent variable models in terms of μ_p and Σ_p is described in several sources (see, e.g., Joreskog, 1977; Muthen, 1983). One common framework is as follows. For a certain population a linear measurement model for a latent variable vector η is specified

(3) $$y = \upsilon + \Lambda \eta + \varepsilon ,$$

where υ and Λ contain measurement intercept and loading (slope) parameters, respectively, and ε denotes a vector of measurement errors. In addition, linear structural equations are specified for η,

(4) $$\eta = \alpha + B\eta + \zeta ,$$

where α and B contain structural regression intercepts and slopes, respectively, and ζ denotes a vector of residuals. With $E(\eta) = \alpha$, $V(\varepsilon) = \Theta$, $V(\zeta) = \Psi$, usual assumptions give the mean and covariance structure for the y vector as

(5) $$\mu = \nu + \Lambda(I - B)^{-1} \alpha,$$

(6) $\Sigma = \Lambda(I - B)^{-1} \Psi (I - B)^{-1'} \Lambda' + \Theta.$

3. A Motivating Example

The example concerns longitudinal observations on mathematics achievement in grades 7-12 collected in the U.S. within the National Longitudinal Study of American Youth (LSAY) (Miller, Suchner, Hoffer, Brown, & Pifer, 1991). Two cohorts were followed, one spanning grades 7-10 and the other grades 10-12. The mathematics curriculum is quite varied in the U.S. and students are likely to show differences in growth as a function of differences in background characteristics such as course taking and gender. The test measures mathematics skills in a number of subtopics including algebra, probability & statistics, geometry, measurement, and arithmetic. Topic-specific subtest scores are of interest, but since there is a rather small number of items within subtopics, there is a need to allow for measurement error in such subscores, for example, by specifying a factor-analytic measurement model.

In order to measure different ability levels, the test items that are administered vary across grades and groups of students within grades. The various test forms do, however, have many items in common so that the various test forms can be equated. Due to the large variation in mathematics achievement, an adaptive testing strategy was employed in the LSAY in order to avoid floor and ceiling effects and to maximize the information obtained on the students' achievement level. Given the performance at the first testing occasion, an easy, medium, or hard test form was chosen for the next grade with possible test form alterations also in subsequent grades. The test forms also differed across grades within difficulty designation. Table 1 shows the different groups of individuals in the youngest cohort taking different sets of tests. It is seen that the adaptive testing strategy gives rise to certain patterns of missing data. Missing data also occurs due to attrition so that not all students have observations for all grades.

As is typical for large-scale educational data, the LSAY data are obtained through multi-stage, complex sampling. A key feature is that about 60 students are randomly sampled within each of about 60 schools. It is well-known that assuming simple random sampling when data have in fact been

Table 1

Missing Data Patterns

		Grade 7		Grade 8				Grade 9				Grade 10		
Sequence[a]	Frequency	Form:	A	Forms:	C	D	E	Forms:	A	B	D	Forms:	A	B
7EEE	503		X		X	--	--		X	--	--		X	--
7EET	97		X		X	--	--		X	--	--		--	X
7EME	242		X		X	--	--		--	--	X		X	--
7EMT	106		X		X	--	--		--	--	X		--	X
7ETT	174		X		X	--	--		--	X	--		--	X
7MME	40		X		--	X	--		--	--	X		X	--
7MMT	113		X		--	X	--		--	--	X		--	X
7MTT	116		X		--	X	--		--	X	--		--	X
7TTT	205		X		--	--	X		--	X	--		--	X
7EE-	210		X		X	--	--		X	--	--		--	--
7EM-	59		X		X	--	--		--	--	X		--	--
7ET-	56		X		X	--	--		--	X	--		--	--
7MM-	14		X		--	X	--		--	--	X		--	--
7MT-	17		X		--	X	--		--	X	--		--	--
7TT-	30		X		--	--	X		--	X	--		--	--
7E--	355		X		X	--	--		--	--	--		--	--
7M--	41		X		--	X	--		--	--	--		--	--
7T--	38		X		--	--	X		--	--	--		--	--
7---	347		X		--	--	--		--	--	--		--	--

Note. X denotes observed data; -- denotes missing data.

[a] 7 denotes 7th grade test (form A)
E denotes easy test (form C in grade 8, A in grades 9, 10)
M denotes medium test (form D in grades 8, 9)
T denotes tough test (form E in grade 8, B in grades 9, 10)

obtained by cluster sampling leads to deflated standard errors of estimates (see, e.g., Skinner, Holt, & Smith, 1989). This effect is often described in terms of the "design effect" (deff), taken as the ratio of the corresponding variance estimates. To illustrate the effect of this cluster sampling feature, intraclass correlations were calculated for a set of achievement variables obtained at the seventh grade. Testlets corresponding to topic-specific sums of items scored right/wrong were used for the following topics (intraclass correlation in parenthesis): algebra (.03), probability & statistics (.15), geometry (.12), measurement (.12), methods (.05), numbers & operations$_1$ (.10), numbers & operations$_2$ (.08), numbers & operations$_3$ (.09), numbers & operations$_4$ (.13), organization (.09). Several intraclass correlations are larger than .10. Using the deff formula for a variance estimate of a mean, 1+ (c-1) ρ for cluster size c and intraclass correlation ρ (Cochran, 1977, p. 242), gives a sizeable design effect of about 7 due to the large cluster size of 60. The intraclass correlations may in fact be deflated since the within-school variance is likely to contain a large amount of measurement error variance (see Muthen, 1991).

4. Modeling of Individual Differences in Growth

For the example discussed in the previous section, consider an achievement score y_{ti} for individual i at time point t where t corresponds to the different grades (t = 0, 1, ..., T, say),

(7) $y_{ti} = \alpha_i + \beta_i\, t + \zeta_{ti}$

In (7), α_i and β_i are individual-specific parameters describing initial level of achievement and rate of learning, while ζ represents a residual. The characteristic feature of this model is that the regression intercepts and slopes are random coefficients that vary over individuals, possibly as a function of individual-specific values of a time-invariant covariate z_i,

(8) $\alpha_i = \alpha + \gamma_\alpha\, z_i + \delta_{\alpha i}$

(9) $\beta_i = \beta + \gamma_\beta\, z_i + \delta_{\beta i}$

Here, α and β represent overall values, γ's are regression parameters, and δ's represent residuals. The residuals for the intercepts and the slopes may be correlated so that the growth rate may be related to initial status. As an example, z may represent participation in enriched or algebra classes, in which case the γ's are likely to be positive. The random intercepts α_i and random slopes β_i may also be estimated for each individual so that an individual-specific growth curve can be derived.

It may be noted that instead of assuming growth that is linear in t, as in (7), any function of t may be used, including functions involving parameters to be estimated, such as logistic growth and exponential decline.

The model implies growth in means and variances as a function of t and z,

$$(10) \quad E(y_{ti} \mid z_i) = \alpha + \gamma_\alpha z_i + (\beta + \gamma_\beta z_i)t$$

$$(11) \quad V(y_{ti} \mid z_i) = \sigma_\alpha^2 + 2\, t\, \sigma_{\alpha\beta} + t^2 \sigma_\beta^2 + \sigma_\zeta^2$$

The model may be extended by adding a time-varying covariate x_{ti} to the growth curve of (7),

$$(12) \quad y_{ti} = \alpha_i + \beta_i t + \gamma_t x_{ti} + \zeta_{ti}$$

In the context of the present achievement example, x_{ti} may represent amount of course work prior to time point t for individual i.

The above growth model can be seen as a model with latent variables. As is clear from (7)–(9), α_i and β_i can be viewed as latent variables instead of random parameters (Muthen, 1991, 1992). Both α_i and β_i are unobserved i.i.d. variables varying across individuals. Because t does not vary over individuals, t can be viewed as a fixed regression parameter for the variable β_i. The model fits into the general framework of equations (3)–(6) letting η contain α_i and β_i.

This type of modeling is an example of the latent curve analysis of Tucker, Meredith, McArdle and others (see, e.g., Meredith & Tisak, 1990). The growth model imposes restrictions on both the mean vector and the covariance matrix

for the observed variables. In this way, both μ and Σ of (1) are used in the estimation. A single population is used.

The structural modeling approach to longitudinal data makes for a very flexible modeling framework. Multiple indicators can be handled so that growth pertains to latent variables without measurement error. In the math achievement example, it is reasonable to assume that the testlets measure a single factor η_{ti}. In this case the factor η_{ti} replaces y_{ti} in (7) and the testlets correspond to multiple indicators y_{tij} as in (3),

(13) $y_{tij} = \nu_j + \lambda_j \eta_{ti} + \varepsilon_{tij}$,

$j = 1, 2, ..., J$, where ν is a measurement intercept parameter, λ is a measurement loading parameter, and ε represents measurement error assumed to be uncorrelated with η and among themselves. Binary and ordered categorical variables can also be handled in this framework (Muthen, 1983, 1992).

5. Modeling of Missing Data

For the motivating example discussed in Section 3, Table 1 showed the pattern of missing data. The missingness was both by design due to the use of adaptive testing and due to attrition. Missing data theory is presented in Little and Rubin (1987) and is discussed in the latent variable context by Allison (1987) and Muthen, Kaplan, and Hollis (1987). Following Muthen et al. (1987), we may modify the measurement model of (3) as

(14) $y^* = \nu + \Lambda \eta + \varepsilon$

(15) $s^* = \Gamma y^* + \delta$

Here, y^* and s^* are sets of r continuous, latent variables assumed to be multivariate normal. The residual vector δ is possibly correlated with η and ε. Using a threshold parameter τ_j, each s^*_{ij} variable defines a probit regression describing the propensity for y^*_{ij} to be observed for individual i.

$$(16)\quad y_{ij} = \begin{cases} y^*_{ij}\text{, if } s^*_{ij} > \tau_j \\ \text{missing, otherwise} \end{cases}$$

Returning to the missing data example of Table 1, consider the first and last missing data patterns. Let the observed test scores in grade 7 be denoted y_1 and the scores of the test sequence E, E, E in grades 8, 9, 10 be denoted y_2. In this way, there is no missingness on y_1 for either pattern, whereas the last pattern has missing data for y_2. Let y_2 contain p variables, define π_i as

$$(17)\quad \Pr(s^*_{i1} \le \tau_1, s^*_{i2} \le \tau_2, ..., s^*_{ip} \le \tau_p) = \pi_i$$

and let ϕ denote multivariate normal densities. The likelihood component for a sample unit in the last missing data pattern is then obtained by integrating over the p latent variables y^*_2 in a truncated normal distribution,

$$(17)\quad \phi(y_{1i})\,\pi_i \int_{-\infty}^{\tau_1} \cdots \int_{-\infty}^{\tau_p} \cdot \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \pi_i^{-1}\,\phi(y^*_2, s^* \mid y_{1i})\, d\, y^*_2\, d\, s^*$$

This gives

$$(18)\quad \phi(y_{1i}) \int_{-\infty}^{\tau_1} \cdots \int_{-\infty}^{\tau_p} \phi(s^* \mid y_{1i})\, d\, s^*$$

The conditional normal density inside the integrals of (18) depends on the specification of the relationship between s* and y* in (14) and (15). Consider the case where conditional on y_{1i}, s* is independent of y^*_2, so that s* is only influenced by y^*_1 in (15). In our example, y^*_1 is observed as y_1. Then the conditional density in (18) does not involve parameters of the latent variable model but only parameters describing how y_1 predicts the missingness on y^*_2. In this case the missing data mechanism is "ignorable" and correct ML estimation of the latent variable model is obtained using only the $\phi(y_{1i})$ term in (18) corresponding to the data that are not missing.

In our example, ignorability for the data that are missing by design holds if the test form for a certain grade is indeed only dependent on the performance on the test in the previous year. Attrition may be predicted by factors that also influence the performance on the tests taken. Missingness by attrition is ignorable if conditional on such factors, the values of the missing test scores are independent of the values of the observed test scores.

Again considering the first and last missing data patterns of Table 1, and assuming ignorability, (18) suggests that the log likelihood may be written as

$$(19)\quad \log L = \sum_{i=1}^{N} \log \phi (y_{1i}) + \sum_{i=1}^{C} \log \phi (y_{2i} \mid y_{1i})$$

where N is the total number of cases in the two patterns and C is the number of individuals that have complete data. The second term on the right hand side of (19) contains the regression parameters, while the first term contains the parameters of the marginal distribution of y_1. As pointed out by Anderson (1957), in the case of an unrestricted model the parameters of these two parts can be estimated separately and the estimates have closed-form expressions. For the case of a latent variable model, the restricted case, a closed-form expression does not, however, exist and the advantage of writing the likelihood in the form of (19) disappears. Muthen et al. (1987) instead proposed the use of the equivalent form

$$(20)\quad \log L = \sum_{i=1}^{C} \log \phi (y_{1i} , y_{2i}) + \sum_{i=C+1}^{N} \log \phi (y_{1i})$$

The two terms of the right hand side of (20) involve two different groups of individuals corresponding to the two different patterns. Equation (20) shows that the standard multiple-group structural modeling fitting function of (1) can be used for the estimation. Under ignorability, a simultaneous analysis of the two groups, using different number of observed variables in the two groups and across-group equality restrictions on common parameters yields ML estimates of the latent variable model parameters. Muthen et al. (1987) describe how to set up this analysis using structural modeling programs and

show how the model can be tested. The approach may be generalized to involve groups corresponding to all the different missing data patterns of Table 1.

6. Modeling of Multilevel Data

The final area to be discussed in terms of latent variable modeling is that of variance components describing data from cluster sampling. In the math achievement example, students were sampled within schools and the intraclass correlation coefficients showed that the degree of dependence among student observations from the same school was quite large. In order for the fitting function of (1) to give proper ML estimates, standard errors of estimates, and chi-square measure of model fit, this deviation from simple random sampling needs to be taken into account. Statistical theory for such situations is described in Skinner, Holt, and Smith (1989). Recently, psychometricians have extended this work to encompass latent variable modeling (see, e.g., McDonald & Goldstein, 1989). For an overview, see Muthen and Satorra (1989), Muthen (1989) and Muthen and Satorra (1991). In this work, parameters are added to those of conventional modeling in order to properly describe the variation due to the different stages of cluster sampling. This has given rise to the name multilevel modeling (see, e.g., Bock, 1989).

The following model describes both the school- and student-level variation. Letting the index g denote school, we may consider the r-dimensional vector of observed scores y_{gi} for individual i and a q-dimensional vector z_g for school g as follows. We may assume g = 1, 2, ..., G independently observed groups with i = 1, 2, ..., N_g individual observations within group g and arrange the data vector for which independent observations are obtained as

(21) $\mathbf{d}_g' = (\mathbf{z}_g', \mathbf{y}_{g1}', \mathbf{y}_{g2}', ..., \mathbf{y}_{gNg}')$,

where we note that the length of $\mathbf{d}_g$ varies across groups. The mean vector and covariance matrix of $\mathbf{d}_g$ are assumed to have the structures

(22) $\mu_{dg}' = [\, \mu_z', 1_{Ng}' \otimes \mu_y' \,]$

$$(23)\quad \Sigma_{d_g} = \begin{bmatrix} \Sigma_{zz} & \text{symmetric} \\ 1_{Ng} \otimes \Sigma_{yz} & I_{Ng} \otimes \Sigma_W + 1_{Ng}\, 1_{Ng}' \otimes \Sigma_B \end{bmatrix}$$

where $\mathbf{I}_{Ng}$ is an identity matrix of dimension $\mathbf{N_g}$, $\mathbf{1}_{Ng}$ is a unit vector of length N_g and the symbol $\otimes$ denotes the Kronecker product.

Assuming multivariate normality of $\mathbf{d}_g$ leads to the minimization of the ML fitting function

$$(24)\quad \sum_{g=1}^{G} \{ \log | \Sigma_{d_g} | + (d_g - \mu_{d_g})' \Sigma_{d_g}^{-1} (d_g - \mu_{d_g}) \}$$

As shown in Muthen (1989, 1990), the expression in (24) may be rewritten in a form that both avoids using parameter arrays involving the number of observations per group and fits in conventional structural equation models. Reducing the summation from G groups to D, corresponding to the number of distinct group sizes, the ML fitting function may be written as

$$(25)\quad \sum_{d}^{D} G_d \{ \ln | \Sigma_{dd} | + \mathrm{tr} [\Sigma_{dd}^{-1} (S_{Bd} + N_d (\bar{v}_d - \mu) (\bar{v}_d - \mu)')] \} +$$

$$+ (N - G) \{ \ln |\Sigma_W| + \mathrm{tr} [\Sigma_W^{-1} S_{PW}] \} ,$$

where d is an index denoting a distinct group size category with group size N_d, G_d denotes the mumber of groups of that size,

$$(26)\quad \Sigma_{dd} = \begin{bmatrix} N_d \Sigma_{zz} & \text{symmetric} \\ N_d \Sigma_{yz} & \Sigma_W + N_d \Sigma_B \end{bmatrix}$$

S_{Bd} denotes a between-group matrix

$$(27)\quad S_{Bd} = N_d\, G_d^{-1} \sum_{k=1}^{G_d} \begin{bmatrix} z_{dk} - \bar{z}_d \\ \bar{y}_{dk} - \bar{y}_d \end{bmatrix} \left[(z_{dk} - \bar{z}_d)' (\bar{y}_{dk} - \bar{y}_d)' \right]$$

$$(28)\quad \bar{v}_d - \mu = \begin{bmatrix} \bar{z}_d - \mu_z \\ \bar{y}_d - \mu_y \end{bmatrix}$$

with $\bar{z}_d$ and $\bar{y}_d$ representing the sample mean vectors in group category d, and S_{PW} is defined as the usual pooled-within sample covariance matrix

$$(29)\quad S_{PW} = (N - G)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} (y_{gi} - \bar{y}_g) (y_{gi} - \bar{y}_g)' .$$

On comparison with (1) it is seen that (25) may be viewed as an analysis of D+1 populations with certain parameter equality constraints across populations.

ML estimation by optimization of (25) is, however, cumbersome with many different group sizes, both in terms of computational work and in terms of input specifications for the software. Muthen (1990) proposed a simpler, ad hoc estimator which gives results close to those of ML, using the fitting function

$$(30)\quad G \{ \ln \begin{vmatrix} c\,\Sigma_{zz} & \text{symmetric} \\ c\,\Sigma_{yz} & \Sigma_W + c\,\Sigma_B \end{vmatrix} + \text{tr} \begin{bmatrix} c\Sigma_{zz} & \text{symmetric} \\ c\,\Sigma_{yz} & \Sigma_W + c\,\Sigma_B \end{bmatrix}^{-1} S_B \} +$$

$$+ (N - G) \{ \ln | \Sigma_W | + \text{tr} [\Sigma_W^{-1} S_{PW}] \} ,$$

where

(31)

$$S_B = (G - 1)^{-1} \begin{bmatrix} c \sum_g (z_g - \bar{z}) (z_g - \bar{z})' & \text{symmetric} \\ c \times G/N \sum_g N_g (\bar{y}_g - \bar{y}) (z_g - \bar{z})' & \sum_g N_g (\bar{y}_g - \bar{y}) (\bar{y}_g - \bar{y})' \end{bmatrix}$$

$$(32)\quad c = [\, N^2 - \sum_{g=1}^{G} N_g^2 \,]\ [\, N\,(G-1)\,]^{-1}$$

and S_{PW} is as before. On comparison with (1) it is seen that (30) corresponds to an analysis with two populations, one for the between part and one for the within part.

For the math achievement test scores of y, a latent variable structure such as in (6) may be formulated for Σ_B and Σ_W, not neccessarily using the same structure. Muthen (1990) discusses different types of models that may be of interest. The within structure of Σ_W would still use a single-factor model since it pertains to the student-level structure. The between structure Σ_B describes across-school variation in math achievement and it is harder to postulate an a priori model for this variation. Experience has shown, however, that a single-factor model often captures the covariation in Σ_B quite well. The school-level variables z_g may be exemplified by indicators of whether or not the school "tracks" the 7th- and 8th-grade math programs. Muthen (1990) gives an example of a latent variable model with z_g variables influencing the between-part of the y variation.

7. Discussion

A thorough analysis of the math achievement example of Section 3 calls for the use of modeling with random coefficients describing individual differences in growth, unobserved variables corresponding to missing data, and variance components describing data from cluster sampling. The previous three sections have described how each of these modeling features may be approached in a general latent variable context using existing structural equation software. The fitting function of (1) is used in all cases, either in one or in several populations using covariance matrix structures and possibly also mean vector structures. In an actual analysis of this data set, the three approaches need to be combined. This analysis will not be carried out here, but it is clear that the use of the fitting function of (1) accomplishes also this complex task.

This paper has made connections between mainstream multivariate statistics and work by psychometricians and other methodologists interested in

latent variable modeling. Viewing the methodology from a general latent variable perspective, points to several interesting extensions of the statistical analyses.

References

Allison, P.D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology, 1987.* San Francisco: Jossey-Bass.

Anderson, T. W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association, 52,* 200-203.

Bock, R.D. (1989). *Multilevel analysis of educational data.* San Diego, CA: Academic Press.

Cochran, W.G. (1977). *Sampling techniques* (3rd ed.). Toronto: John Wiley & Sons.

Joreskog, K.G. (1977). Structural equation models in the social sciences: Specification, estimation and testing. In P.R. Krishnaiah (Ed.), *Applications of statistics.* Amsterdam: North-Holland.

Little, R.J., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley & Sons.

McDonald, R.P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42,* 215-232.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107-122.

Miller, J. D., Suchner, R. W., Hoffer, T., Brown, K. G., & Pifer, L. (1991). *LSAY codebook: Student, parent, and teacher data for cohort one for longitudinal years one, two, and three (1987-1990).* De Kalb: Northern Illinois University.

Muthen, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22,* 43-65.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthen, B. (1989). Latent variable modeling in heterogeneous populations. Presidential address to the Psychometric Society, July, 1989. *Psychometrika, 54,* 557-585.

Muthen, B. (1990, May). *Mean and covariance structure analysis of hierarchical data* (UCLA Statistics Series #62). Los Angeles: University of California.

Muthen, B. (1991a). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.

Muthen, B. (1991b). Analysis of longitudinal data using latent variable models with varying parameters. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 1-17). Washington DC: American Psychological Association.

Muthen, B. (1992). *Random coefficient growth modeling in a structural equations framework* (Tech. Rep.). Los Angeles: University of California.

Muthen, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 42*, 431-462.

Muthen, B., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 87-99). San Diego: Academic Press.

Muthen, B., & Satorra, A. (1991). *Complex sample data in structural equation modeling*. Los Angeles: University of California.

Skinner, C.J., Holt, D., & Smith, T.M.F. (1989). *Analysis of complex surveys*. Chichester: John Wiley & Sons.